AD________________

Award Number: DAMD17-98-1-8044

TITLE: A Computer-Based Decision Support System for Breast Cancer Diagnosis

PRINCIPAL INVESTIGATOR: Lan Luo

CONTRACTING ORGANIZATION: The Catholic University of America
Washington, DC 20064

REPORT DATE: September 1999

TYPE OF REPORT: Annual Summary

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for public release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)

2. REPORT DATE
September 1999

3. REPORT TYPE AND DATES COVERED
Annual Summary (1 Sep 98 - 31 Aug 99)

4. TITLE AND SUBTITLE
A Computer-Based Decision Support System for Breast Cancer Diagnosis

5. FUNDING NUMBERS
DAMD17-98-1-8044

6. AUTHOR(S)
Lan Luo

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
The Catholic University of America;
Washington, DC 20064

E-MAIL:
zwang@pluto.ee.cua.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***

The optimal goal of this project is to develop decision support systems for breast cancer diagnosis, treatment option, prognosis, and risk prediction. The primary goal for the first year is to develop visual presentation methods for the consultation system. We have developed an automated and intelligent procedure for generating the hierarchy of minimax entropy models and principal component visualization spaces for improving data explanation. The proposed model is both statistically principled and visually effective at revealing all of the interesting aspects of the data set. The methods involve multiple use of standard finite normal mixture models and probabilistic principal component projections, whose parameters are estimated using expectation-maximization algorithm and principal component neural networks under the information theoretic criteria. The strategy is that top level model and projection should explain the entire data set, best revealing the presence of clusters and their relationships, while lower level models and projections should display internal structure in individual cluster, such as the presence of subclusters and attribute trends. We have demonstrated the principle of this approach on two three-dimensional synthetic data sets, and we then applied the method to the visual explanation in computer-aided diagnosis for breast cancer detection from digital mammograms.

14. SUBJECT TERMS
Breast Cancer Diagnosis, Breast Cancer Patient Database, Decision Support System, Computer-Aided Diagnosis, Artificial Intelligence

15. NUMBER OF PAGES
11

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT
Unclassified

18. SECURITY CLASSIFICATION OF THIS PAGE
Unclassified

19. SECURITY CLASSIFICATION OF ABSTRACT
Unclassified

20. LIMITATION OF ABSTRACT
Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

Z.W. Where copyrighted material is quoted, permission has been obtained to use such material.

Z.W. Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

Z.W. Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

N/A In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and use of Laboratory Animals of the Institute of Laboratory Resources, national Research Council (NIH Publication No. 86-23, Revised 1985).

X For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

N/A In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

N/A In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

N/A In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

For L. Luo

Zuyi Wang 9/27/99

PI - Signature Date

TABLE OF CONTENTS

1. Introduction

The goal of this project is to develop decision support system for breast cancer diagnosis, treatment option, prognosis, and risk prediction. This system is desired to function as a consultation system for both doctors and patients. This predoctoral research project is focusing on the development of advanced image pattern analysis in diagnostic imaging and information integration methodology. The **specific aims** of this research project are: (1) image pattern analysis of breast tissue in mammography using both computational features and BI-RADS features provided by radiologist for the prediction of malignancy associated with masses; (2) development of visual presentation methods for radiologists' use in the consultation system; (3) performing a pre-clinical test through an ROC analysis. The **clinical goal** of this consultation system is to provide scientific tools for doctors to have electronic magnification views, to perform feature analysis of suspected mammographic patterns, to access a large database and investigate clinically similar cases, and to visually inspect the features of a case in various statistical distribution using graphic displays. The **primary objective** of the Phase I research is to develop visual explanation tools with an interactive presentation of cases and statistics, features and patterns, evidence and narrative.

2. Overview of Professional Training

2.1 Experiment Observation and Case Study

During the three months in summer 1999 working at Imaging Science and Information Systems Research Center, Georgetown University Medical Center, I discussed with clinical mentor Dr. Matthew Freedman about medical issue of breast cancer, for example, the development of breast cancer, various lesions found in breast and BI-RADS features and their importance for analysis of mammogram, etc. I have also observed the some mammography procedures in the hospital and read some mammograms with radiologists. All these experience helped me better understand the breast cancer and mammography, and gave me better ideas of what type of diagnosis consultation system we should develop to provide the assistance that radiologists need. I worked with technical mentor Dr. Ben Lo on the development of breast cancer database, and I discussed with him about technical problems we confront for feature selection and extraction. I learnt more engineering knowledge and computer skills for my future research from our mentors.

2.2 Book Reading and Literature Survey

By reading *Breast Imaging* by D. B. Kopans, I learn more about breast cancer and breast imaging, e.g. breast anatomy, stages of breast cancer, principles of X-ray mammography, lesion features and mammography interpretation, etc. I read more technical papers about data visualization and feature extraction, some of the key papers are: *A Hierarchical Latent Variable Model for Data Visualization* by Bishop and Tipping; *Mixtures of Probabilistic Principal Component Analyzers* by Tipping and Bishop; *Maximum Likelihood Neural Networks for Sensor Fusion and Adaptive Classification* by Perlovsky and McManus; *Computerized Analysis of Mammographic Microcalcifications in Morphological and Texture Feature Spaces* by Chan and Sahiner et. al.; *Computerized Characterization of Masses on Mammograms: the rubber Band Straightening Transform and Texture Analysis* by Sahiner and Chan et. al. Through paper reading I learnt about other related previous research so that I am able to look deeply into some existing problems and set this as a start of our research.

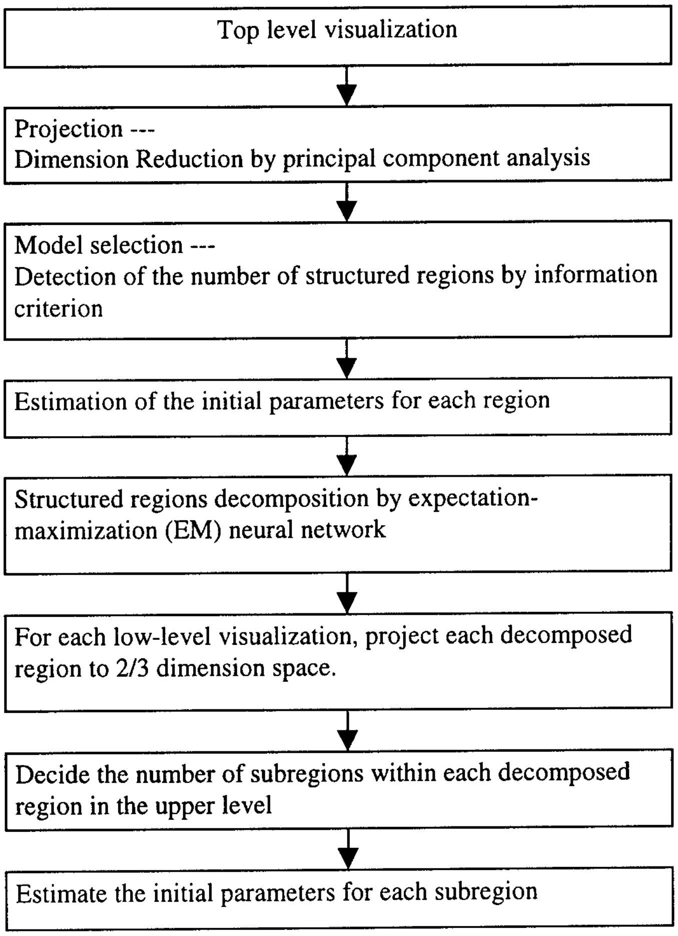
3. Overview of Project Development

3.1 Background

The major research efforts in the first year have been contributed to development and improvement of the visual explanation tools for multivariate data mining. Most of existing visualization algorithms aim to find to projection from the data space down to a visually perceivable rendering space. However, most of such algorithms are based on a projection of the data onto a two-dimensional visualization space, which is often proven to be inadequate in the case of more complex and high dimensional data sets. For example, a single projection of the data onto a visualization space may not be able to capture all of the interesting and important aspects of the data set. The shortcoming of the existing algorithms motivates our consideration of a hierarchical visualization paradigm involving multiple statistical models and visualization spaces. We introduced a hierarchical visualization algorithm that allows the complete data set to be visualized at the top level, with clusters and subclusters of data points visualized at deeper levels.

3.2 Features of Hierarchical Visualization Model

We have developed new techniques for data visualization and interpretation using multiple finite normal mixture models and hierarchical visualization spaces, following the steps showing in Figure 1. The strategy is that the top-level model and projection should explain the entire data set and best reveal the presence of clusters and their relationships, while lower-level models and projections should display internal structure of individual clusters, such as the presence of subclusters, which might not be apparent in the high-level models and projection.


Top level visualization
Projection ---
Dimension Reduction by principal component analysis
Model selection ---
Detection of the number of structured regions by information criterion
Estimation of the initial parameters for each region
Structured regions decomposition by expectation-maximization (EM) neural network
For each low-level visualization, project each decomposed region to 2/3 dimension space.
Decide the number of subregions within each decomposed region in the upper level
Estimate the initial parameters for each subregion

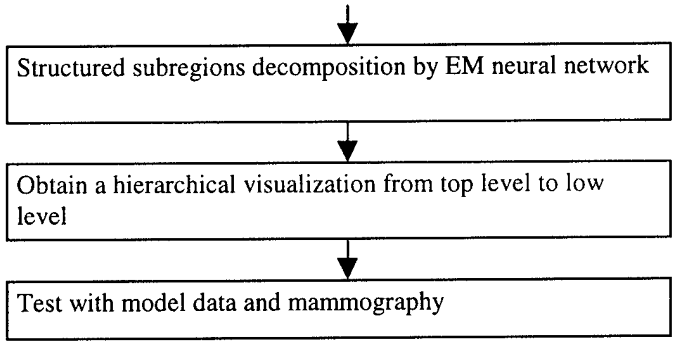

Structured subregions decomposition by EM neural network
Obtain a hierarchical visualization from top level to low level
Test with model data and mammography

Figure 1. Development of Hierarchical Visualization Algorithm

Based on the concept of combining minimax entropy modeling and principal component analysis, our algorithm optimizes structure decomposition and dimensionality reduction. The particular advantages of our algorithm are:

1. At each level, a probabilistic principle component analysis is performed to project the softly partitioned data space down to a desired two-dimensional visualization space, leading to an optimal dimensionality reduction that allows best separation and visualization of local clusters;
2. Learning from the data directly, minimax entropy principle is used to select model structures and estimate its parameter values, where the soft partitioning of the data set results in a standard finite normal mixture model with minimum conditional bias and variance;
3. By performing principal component analysis (PAC) and minimax entropy modeling alternatively, a complete hierarchy of complementary projections and refined models can be generated automatically, corresponding to a statistical description best fit to the data.

Several major differences between our work and most of previous related research, which highlight the significance of the new technique, include:

1. Structure decomposition and dimensionality reduction are two separated but complementary operations, where criterion used to optimize dimensionality reduction is the separation of clusters rather than maximum likelihood;
2. The number of subclusters inside each cluster at each level is determined by a model selection procedure using the information theoretic criteria, which allows the algorithm to automatically determine whether a further split of a subspace should be continued or terminated in completing the whole hierarchy;
3. A probabilistic adaptive principal component extraction algorithm has been developed to estimate the desired number of principal axes. This method is very computationally efficient when the dimension of raw data set is high.
4. Our model defines a probability distribution in data space that naturally induces a corresponding distribution in projection space through Radon transform, which

permits an independent procedure in determining values for intrinsic model parameters without concurrent estimation of projection mapping matrix.

3.3 Experiments and Applications

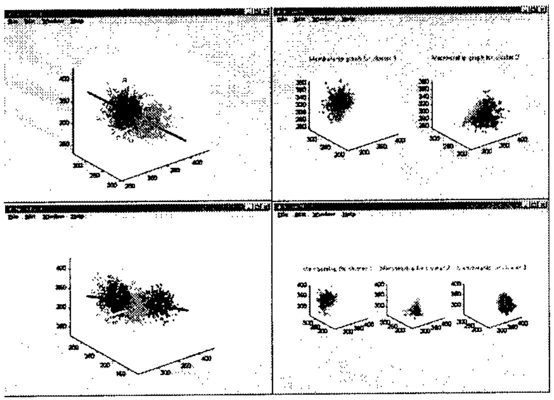


Figure 2. Soft clustering data set, top and bottom figures showing two independent experiments.

We have applied our model to several examples, including simple and complex synthetic data, and breast cancer detection in computer-aided diagnosis (CAD). As a simple example, shown in Figure 2 (top figures), a synthetic data set with a mixture of two Gaussians in a three-dimensional space is considered. In order to illustrate global and local principal axes, these two cloud-like clusters are designed to be relatively overlapped. To explore the data, we first performed a single global PCA to extract the global principal axis. Two information theoretic criteria have clearly suggested the presence of two distinct clusters in the data set. The user selects two initial cluster centers and then EM algorithm is applied to conduct a soft clustering of data points. This leads to a mixture of two probabilistic principal component subspaces whose principal axes are separately extracted. The same operation has been done to a more complex data set consisting of a mixture of three Gaussians, also shown in Figure 2 (bottom figures), two of cloud-like clusters are well separated while the third is spaced in between. A second level visual space is generated with a mixture of two local PCA subspaces where the black line indicates the global principal axis. At the third level data modeling, a hierarchical model is adopted, which illustrates that there are indeed total three clusters in the data set.

We have also applied the probabilistic modular neural network (PMNN) to the so-called "M+1 classes" problem, in which the disease pattern under testing could be either from one of the M class, or from some other unknown class. We observed consistent and significant results compared with pure Bayesian decision. Using the ORL standard database, it has been shown that the correct detection rate is increased from 70% to 90%.

In the application of our algorithm to real mammograms, we used the classifier to distinguish true mass from false mass based on the features extracted from the suspected regions. 150 mammograms were selected from the mammogram database, each of them contains at least one mass case of varying size and location. The areas of suspicious masses were identified following the procedure with biopsy proven result. 50 mammograms with biopsy proven masses were selected from the data set for training. The mammogram set used for testing contained 46 single view mammograms (23 normal cases and 23 with biopsy proven masses). We selected a 3-D feature space consisting of compactness I, compactness II and entropy. A training feature vector set was constructed from 50 true mass region of interests (ROI) and 50 false mass ROIs. It is shown that the recognition rate with compactness I is more reliable than compactness II.

We have conducted a preliminary study to evaluate the performance of the algorithm in real case detection, in which 6 – 15 suspected masses per mammogram were detected and required for the clinical decision making. We found that the classifier could reduce the number of suspicious masses with a sensitivity of 84% at a specificity of 82% based on the database containing 46 mammograms. The results indicated that stellate mass lesion was correctly detected. These preliminary tests showed that knowledge database mapping by PMNN and information theoretic criteria, based on the most representative common database, is promising approach.

4. Conclusion

We have had successful progress in both training and research. Through discussion with clinical mentor and radiologists and observation of mammogram reading, we learnt a lot of relevant medical knowledge and radiologists' needs so that our research is led to the right direction to develop a useful and practical diagnosis consultation system. The better understanding of digital mammogram and patient database, as well as engineering knowledge and computer skill, also enable us to identify and solve the problems we encountered in our work.

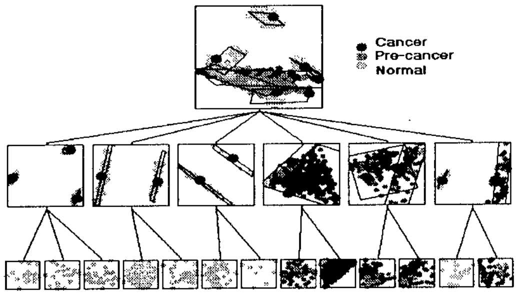


Figure 3. Hierarchical visualization model for high dimensional data set exploration

We have developed the hierarchical data visualization model for improving data explanation, which is both statistically principled and visually effective at revealing all interesting aspects of data set. This method, as illustrated by the well-planned simulations, shown in Figure 3, can be very capable of revealing hidden structure in data set. Since the models are determined by minimax entropy, and this criterion not only can estimate the parameters but also select the structure, this approach allows a self-consistent fitting of the whole hierarchical tree. Maximum separation of clusters in turn optimizes other crucial operations, such as model selection and parameter initialization. While the optimality of these techniques are often highly data-dependent, we would expect the hierarchical visualization model to be very effective for data visualization and exploration in many applications.

Appendix

Key Research Accomplishments

- We have constructed the feature knowledge database from all the suspicious mass sites localized by the enhanced segmentation using a mathematical feature extraction procedure.
- We have developed the hierarchical visualization model for multivariate data mining, which is statistically principled and visually effective. This method has been demonstrated to be very capable of revealing hidden structure in data sets.
- We have further invented a visual explanation tool for the decision making as a clinical support, based on the interactive visualization hierarchy through the probabilistic principal component projection of the knowledge database and the localized optimal displays of the retrieved raw data.
- We have developed a prototype system and pilot tested to demonstrate the applicability of the framework to mammographic mass detection.

Reportable Outcomes

1. Wang,Y., Luo, L., Li, H. and Freedman, M.T., "Hierarchical Minimax Entropy Modeling and Probabilistic Principal Component Visualization for Data Exploration", *Proc. SPIE,* Vol. 3658, pp. 108-116, 1999.
2. Luo, L., Wang, Y. and Kung, S.Y., "Hierarchy of Probabilistic Principal Component Subspaces for Data Mining", *Proc. IEEE Workshop on Neural Networks for Signal Processing,* Wisconsin, August 1999.
3. Wang, Y., Luo, L., Freedman, M.T. and Kung, S.Y., "Probabilistic Principal Component Subspaces: A Hierarchical Minimax Entropy Model for Data Visualization", submitted to *IEEE Trans. Neural Networks, 1999.*
4. Wang, Y., Luo, L., Freedman, M.T. and Kung, S.Y., "Hierarchical Probabilistic Principal Component Subspaces for Data Visualization", *Proc. Intl. Conf. Neural Networks,* Washington, DC, July 1999.